

On the computational complexity of stochastic controller optimization in POMDPs

Nikos Vlassis* Michael L. Littman† David Barber‡

July 18, 2011

Abstract

We show that the problem of finding an optimal stochastic ‘blind’ controller in a Markov decision process is an NP-hard problem. The corresponding decision problem is NP-hard, in PSPACE, and SQRT-SUM-hard, hence placing it in NP would imply a breakthrough in long-standing open problems in computer science. Our optimization result establishes that the more general problem of stochastic controller optimization in POMDPs is also NP-hard. Nonetheless, we outline a special case that is solvable to arbitrary accuracy in polynomial time via semidefinite or second-order cone programming.

Keywords: Partially observable Markov decision process, stochastic controller, bilinear program, computational complexity, Motzkin-Straus theorem, sum-of-square-roots problem, matrix fractional program, semidefinite programming.

1 Introduction

Partially observable Markov decision processes (POMDPs) have proven to be a valuable conceptual tool for problems throughout AI, including reinforcement learning (Chrisman, 1992), planning under uncertainty (Kaelbling et al., 1998), and multiagent coordination (Bernstein et al., 2005). Briefly, a POMDP consists of a Markov process over a set of states. The decision maker is unable to perceive its current state directly, but must infer it based on indirect observations. An important problem in this area is deciding how to select actions to minimize cost given the state uncertainty. Unfortunately, this problem is extremely challenging (Papadimitriou and Tsitsiklis, 1987; Mundhenk et al., 2000). In fact, the exact problem is unsolvable in the general case (Madani et al., 1999).

An alternative to finding optimal policies for POMDPs is to find low cost *controllers*—mappings from observation histories to actions (Sondik, 1971; Platzman, 1981). A restricted space of controllers can, in principle, be considerably easier to search than the space of all possible policies (Littman et al., 1998; Hansen, 1998; Meuleau et al., 1999). Various methods for controller optimization in POMDPs have been proposed in the literature, both for stochastic as well as for deterministic controllers: exhaustive search (Smith, 1971), branch

*Luxembourg Centre for Systems Biomedicine, Univ. of Luxembourg (nikos.vlassis@uni.lu)

†Department of Computer Science, Rutgers University (mlittman@cs.rutgers.edu)

‡Department of Computer Science, University College London (d.barber@cs.ucl.ac.uk)

and bound (Hastings and Sadjadi, 1979; Littman, 1994), local search (Poupart and Boutilier, 2004; Serin and Kulkarni, 2005), sequential quadratic programming (Amato et al., 2007), or the EM algorithm (Toussaint et al., 2011).

A variety of complexity results are known for the problem of controller optimization in POMDPs. Most versions are known to be hard for classes that are believed to be above P (Papadimitriou and Tsitsiklis, 1987; Mundhenk et al., 2000). The computational decision problem is: Given a restriction on the controller and a target cost, can the target cost be achieved by a controller in the class? Below, we consider several such controller classes.

Deterministic time/history-dependent controller Such a controller selects an action based on the current time period and/or the history of previous actions and observations. The problem is NP-complete or PSPACE-complete (Papadimitriou and Tsitsiklis, 1987; Mundhenk et al., 2000). In the remaining classes below we assume stationary controllers.

Deterministic controller of polynomial size Such a controller is represented by a graph in which nodes are labeled with actions and edges are labeled with observations. A deterministic controller can approximate the optimal policy for any POMDP. The problem is in NP in that we can guess a controller of the right size, then see if achieves no more than the target cost by solving a system of linear equations. It is NP-hard even for the ‘easier’ completely observable version (Littman et al., 1998).

Stochastic controller of polynomial size This class extends deterministic controllers by allowing a probability distribution over actions at each node. There are POMDPs for which a stochastic controller of a given size can outperform any deterministic controller of the same size (Singh et al., 1994). In this paper we show that this problem is NP-hard, in PSPACE, and SQRT-SUM-hard, hence showing it lies in NP would imply breakthroughs in long-standing open problems (Allender et al., 2009; Etessami and Yannakakis, 2010).

Deterministic memoryless controller A memoryless controller chooses an action based on the most recent observation only. These controllers are a special case of deterministic controllers with polynomial size as they can be represented as a graph with one node per observation. The problem has been shown to be NP-complete (Littman, 1994).

Stochastic memoryless controller These controllers are defined by a probability distribution over actions for each observation. They can be considerably more effective than the corresponding deterministic memoryless controllers. They are a generalization of the blind controllers we consider in this paper, and it follows from our results that the problem is NP-hard, in PSPACE, and SQRT-SUM-hard.

Deterministic blind controller A blind controller for a POMDP is equivalent to a memoryless controller for an unobserved MDP. A deterministic blind controller consists of a single action that is applied (blindly) regardless of the observation history. It is straightforward to evaluate a deterministic blind controller—simply drop all actions but one from the POMDP and evaluate the resulting Markov chain. Thus, the decision problem for deterministic blind controllers is trivially in P as an algorithm can simply try each action to see which is best.

Stochastic blind controller Such a controller is a probability distribution over actions to be applied repeatedly at every timestep. This is the class of controllers we consider in this paper. Again, the added power of stochasticity allows for much more effective policies to be constructed. However, as we show in the remainder of this paper, the added power comes with a very high cost. The decision problem is NP-hard, in PSPACE, and SQRT-SUM-hard.

2 MDPs and Blind Controllers

We consider a discounted, with discount factor $\gamma < 1$, infinite-horizon Markov decision process (MDP) characterized by n states and k actions, state-action costs (negative rewards) c_{sa} , and starting distribution (μ_s) with $\mu_s \geq 0$ and $\sum_{s=1}^n \mu_s = 1$. Let $p(\bar{s}|s, a)$ denote the probability to transition to state \bar{s} when action a is taken at state s . The following linear program (LP) can be used to find an optimal policy for the MDP:

$$\begin{aligned} \min_{x_{sa}} \quad & \sum_{sa} x_{sa} c_{sa}, \\ \text{s.t.} \quad & \sum_a x_{\bar{s}a} = (1 - \gamma)\mu_{\bar{s}} + \gamma \sum_{sa} p(\bar{s}|s, a) x_{sa} \quad \forall \bar{s}, \quad x_{sa} \geq 0 \quad \forall s, a, \end{aligned} \tag{1}$$

where x_{sa} denotes occupancy distribution over state-action pairs, and the constraints are the Bellman flow (probability mass) constraints. From an optimal occupancy x_{sa}^* , we can compute an optimal stationary and deterministic policy that maps states to actions (Puterman, 1994).

We consider now the case where we constrain the class of allowed policies to stochastic ‘blind’ controllers in which the controller cannot observe or remember anything (state, action, or time). Instead, the controller simply randomizes over actions using the same distribution $\boldsymbol{\pi} = (\pi_a)$ at each time step, where $\boldsymbol{\pi} \in \Delta$ and $\Delta = \{\boldsymbol{\pi} : \boldsymbol{\pi} \geq 0, \sum_{a=1}^k \pi_a = 1\}$ is the standard probability simplex. Note that, contrary to standard MDP policies, a blind controller $\boldsymbol{\pi}$ is *not* a function of state. (The related notion of a memoryless controller is a function of POMDP observations, but still not of state.) Explicitly encoding the controller

parametrization in (1) gives:

$$\begin{aligned} \min_{\mathbf{x} \geq 0, \boldsymbol{\pi} \in \Delta} \quad & \sum_{sa} x_s \pi_a c_{sa}, \\ \text{s.t.} \quad & x_{\bar{s}} = (1 - \gamma) \mu_{\bar{s}} + \gamma \sum_a \pi_a \sum_s p(\bar{s}|s, a) x_s \quad \forall \bar{s}, \end{aligned} \quad (2)$$

where $\mathbf{x} = (x_s)$ is an occupancy distribution over states, with $\mathbf{x} \geq 0$. When viewed as a function of both \mathbf{x} and $\boldsymbol{\pi}$, the above constitutes a jointly constrained bilinear program that is in general nonconvex in $(\mathbf{x}, \boldsymbol{\pi})$ (Al-Khayyal and Falk, 1983).

Bilinear programs are known to be NP-hard to solve to global optimality in general, but could there be some special structure in (2) that renders that particular program tractable? In the next section, we answer this question for the case where the MDP costs c_{sa} depend nontrivially on both states and actions, in which case we show that finding an optimal stochastic blind controller is an NP-hard problem.

3 NP-hardness Result

Let $\mathbf{C} = (c_{sa})$ be the $n \times k$ matrix containing all state-action costs, and $\boldsymbol{\mu} = (\mu_s)$ be the $n \times 1$ starting distribution vector. The decision version of our problem, henceforth called the STOCHASTIC-BLIND-POLICY problem, asks, for a given MDP with discount factor $\gamma < 1$ and a given target value r , whether there exists a stochastic blind controller $\boldsymbol{\pi}$ that achieves $J(\boldsymbol{\pi}) \leq r$, where $J(\boldsymbol{\pi}) = \mathbf{x}^\top \mathbf{C} \boldsymbol{\pi}$ is the value of controller $\boldsymbol{\pi}$ in (2) when the $n \times 1$ occupancy vector \mathbf{x} is defined via the Bellman constraints in (2).

Theorem 1. *The STOCHASTIC-BLIND-POLICY problem is NP-hard.*

Proof. We reduce from the INDEPENDENT-SET problem. This problem asks, for a given (undirected and with no self-loops) graph $G = (V, E)$ and a positive integer $j \leq |V|$, whether G contains an independent set V' having $|V'| \geq j$. This problem is NP-complete, even when restricted to cubic planar graphs (Garey and Johnson, 1979).

Let \mathbf{G} be the $n \times n$ (symmetric, 0–1) adjacency matrix of an input cubic graph G . The reduction constructs an MDP with n states and n actions, uniform starting distribution $\boldsymbol{\mu}$, cost matrix $\mathbf{C} = \frac{1}{\gamma}(\mathbf{G} + \mathbf{I})$ where \mathbf{I} is the identity matrix, and deterministic transitions $p(\bar{s}|s, a) = 1$ if $\bar{s} = a$ and 0 otherwise (where the action variable a can be viewed as indexing the state space). Since the transitions $p(\bar{s}|s, a)$ are independent of s , the occupancy vector in (2) reduces to $\mathbf{x} = (1 - \gamma)\boldsymbol{\mu} + \gamma\boldsymbol{\pi}$, and the value function becomes the quadratic

$$J(\boldsymbol{\pi}) = \frac{4(1 - \gamma)}{n\gamma} + \boldsymbol{\pi}^\top (\mathbf{G} + \mathbf{I}) \boldsymbol{\pi}, \quad (3)$$

where we used the fact that the input graph is cubic (each node has degree three) and $\boldsymbol{\mu}$ is uniform. The Motzkin-Straus theorem (Motzkin and Straus, 1965) states that

$$\frac{1}{\alpha(G)} = \min_{\boldsymbol{\pi} \in \Delta} \boldsymbol{\pi}^\top (\mathbf{G} + \mathbf{I}) \boldsymbol{\pi}, \quad (4)$$

where $\alpha(G)$ is the size of the maximum independent set (the stability number) of the graph. Let the target value be $r = \frac{1}{j} + \frac{4(1-\gamma)}{n\gamma}$. Then, $J(\boldsymbol{\pi}) \leq r$ is equivalent to $\boldsymbol{\pi}^\top (\mathbf{G} + \mathbf{I}) \boldsymbol{\pi} \leq \frac{1}{j}$, and hence from (4) follows that the existence of a vector $\boldsymbol{\pi}$ that satisfies $J(\boldsymbol{\pi}) \leq r$ would imply $\frac{1}{\alpha(G)} \leq \frac{1}{j}$, and hence $\alpha(G) \geq j$, or, in other words, $|V'| \geq j$ for some V' . \square

4 On the Complexity Upper Bound

Our STOCHASTIC-BLIND-POLICY problem is contained in PSPACE, as it can be expressed as a system of polynomial inequalities—any such system is known to be solvable in PSPACE (Canny, 1988). But, is there a tighter upper bound?

We will attempt to address this question indirectly, by establishing a connection between the STOCHASTIC-BLIND-POLICY problem and the SQRT-SUM problem. The SQRT-SUM problem asks, for a given list of integers c_1, \dots, c_n and an integer d , whether $\sum_{i=1}^n \sqrt{c_i} \leq d$. The problem is conjectured to lie in P, but it is not even known to lie in NP. The difficulty of obtaining an exact complexity for this problem has been recognized for at least 35 years (Garey et al., 1976). Allender et al. (2009) showed that SQRT-SUM lies in the 4th level of the Counting Hierarchy, and Etessami and Yannakakis (2010) showed that SQRT-SUM reduces to the problem of approximating 3-player Nash equilibria. Here, we show that STOCHASTIC-BLIND-POLICY is at least as hard as SQRT-SUM, hence a result that would place STOCHASTIC-BLIND-POLICY in NP would resolve several open problems in computer science (Allender et al., 2009; Etessami and Yannakakis, 2010).

Theorem 2. *The STOCHASTIC-BLIND-POLICY problem is SQRT-SUM-hard.*

Proof. Let c_1, \dots, c_n and d be the inputs of SQRT-SUM. The reduction constructs an MDP with $n+1$ states and n actions, where the $(n+1)$ th state is absorbing (self-looping). The starting probabilities are $\mu_i = \frac{1}{n}$ for states $i = 1, \dots, n$ and $\mu_{n+1} = 0$, and the costs depend only on state and are given by the inputs c_i for states $i = 1, \dots, n$ and $c_{n+1} = 0$. From each state $i = 1, \dots, n$, the i th action deterministically stays at state i while all other actions deterministically transition to the absorbing state $n+1$.

For each state $i = 1, \dots, n$, the Bellman occupancy constraint reads $x_i = (1-\gamma)/n + \gamma\pi_i x_i$, and the value function becomes:

$$J(\boldsymbol{\pi}) = \sum_{i=1}^n c_i x_i = \frac{1-\gamma}{n} \sum_{i=1}^n \frac{c_i}{1-\gamma\pi_i}. \quad (5)$$

Differentiating J with respect to $\boldsymbol{\pi}$ after introducing a Lagrange multiplier λ for the constraint $\sum_i \pi_i = 1$, and setting to zero, gives an equation that involves λ and π_i . We can eliminate λ from that equation by solving for each π_i and then using $\sum_i \pi_i = 1$, resulting in an optimal multiplier

$$\lambda^* = \frac{\gamma(1-\gamma)}{n(n-\gamma)^2} \left(\sum_{i=1}^n \sqrt{c_i} \right)^2. \quad (6)$$

Substituting in (5) the (irrational) $\boldsymbol{\pi}^*$ corresponding to λ^* we get the optimal value:

$$J^* = \frac{1-\gamma}{n(n-\gamma)} \left(\sum_{i=1}^n \sqrt{c_i} \right)^2. \quad (7)$$

The STOCHASTIC-BLIND-POLICY question of whether there exists a stochastic blind controller $\boldsymbol{\pi}$ with value $J(\boldsymbol{\pi}) \leq r$ is clearly equivalent to the question whether $J^* \leq r$. By choosing $r = \frac{(1-\gamma)d^2}{n(n-\gamma)}$, we see from (7) that the condition $J^* \leq r$ is equivalent to $\sum_{i=1}^n \sqrt{c_i} \leq d$, and the reduction is complete. \square

5 A Special Case that is in P

We outline here a special case that is solvable to arbitrary accuracy in polynomial time via semidefinite or second-order cone programming, and a variant in which the exact optimal solution can be computed in polynomial time.

For each action a , let \mathbf{P}_a denote the corresponding MDP transition matrix, $\mathbf{P}_a(\bar{s}, s) = p(\bar{s}|s, a)$. The special case assumes that each matrix \mathbf{P}_a is symmetric (and therefore doubly stochastic). The bilinear program (2) then reads:

$$\min_{\boldsymbol{\pi} \in \Delta} (1-\gamma) \boldsymbol{\pi}^\top \mathbf{C}^\top (\mathbf{I} - \gamma \mathbf{M}_\boldsymbol{\pi})^{-1} \boldsymbol{\mu}, \quad (8)$$

where $\mathbf{M}_\boldsymbol{\pi} = \sum_a \pi_a \mathbf{P}_a$.

Lemma 1. *For any $\boldsymbol{\pi}$, the matrix $\mathbf{I} - \gamma \mathbf{M}_\boldsymbol{\pi}$ is positive definite.*

Proof. Since each matrix \mathbf{P}_a is symmetric and stochastic, all its eigenvalues are real and satisfy $\lambda(\mathbf{P}_a) \leq 1$. Hence, the eigenvalues of $\mathbf{I} - \gamma \mathbf{P}_a$ are also real and satisfy $\lambda(\mathbf{I} - \gamma \mathbf{P}_a) = 1 - \gamma \lambda(\mathbf{P}_a) > 0$ because $\gamma < 1$. Therefore, $\mathbf{I} - \gamma \mathbf{P}_a$ is a positive definite matrix, and so must be the matrix $\mathbf{I} - \gamma \mathbf{M}_\boldsymbol{\pi}$ as it can be written as the convex combination (over $\boldsymbol{\pi}$) of positive definite matrices. \square

If we constrain the feasible region to those $\boldsymbol{\pi}$ for which $\mathbf{C}\boldsymbol{\pi} = \kappa \boldsymbol{\mu}$, with $\kappa \in \mathbb{R}$, then we can formulate the program (8) as a *matrix fractional program*, which, by taking epigraph and a Schur complement, and using Lemma 1, can be expressed as a convex program involving a linear matrix inequality and linear constraints:

$$\begin{aligned} \min_{t \in \mathbb{R}, \kappa \in \mathbb{R}, \boldsymbol{\pi} \in \Delta} \quad & t \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{I} - \gamma \mathbf{M}_\boldsymbol{\pi} & \boldsymbol{\mu} \\ \boldsymbol{\mu}^\top & t \end{bmatrix} \succeq 0, \quad \mathbf{M}_\boldsymbol{\pi} = \sum_a \pi_a \mathbf{P}_a, \quad \mathbf{C}\boldsymbol{\pi} = \kappa \boldsymbol{\mu}, \end{aligned} \quad (9)$$

which can be solved efficiently to arbitrary accuracy by semidefinite programming or second-order cone programming (Boyd and Vandenberghe, 2004).

If we further assume that the costs are nonpositive and satisfy $\mathbf{C} = -\kappa\boldsymbol{\mu}\mathbf{1}^\top$, with $\kappa > 0$, then (8) becomes a minimization of a concave function over the probability simplex, hence its optima will appear in a corner of the simplex and the optimal controller will be deterministic. Since there are only k deterministic controllers, evaluating each of them and selecting the optimal one takes $O(kn^3)$ operations.

6 Conclusions

In response to the computational intractability of searching for optimal policies in POMDPs, many researchers have turned to finite-state controllers as a more tractable alternative. We have provided here a computational characterization of exactly solving problems in the class of stochastic controllers, showing that (1) they are NP-hard, (2) they are in PSPACE, and (3) they are SQRT-SUM-hard, hence showing membership in NP would resolve long-standing open problems.

We note that our NP-hardness proof relies on the assumption that the costs c_{sa} are nondegenerate functions of both state and action. We have been unable to extend the NP-hardness proof to the case where the costs are functions of state only. Although the proof of SQRT-SUM-hardness employs such costs, no hardness result above polynomial time is known for SQRT-SUM, leaving the complexity of the case of state-only-dependent costs of the stochastic blind controller problem open.

In this work, we only addressed the complexity of the decision problem for the discounted infinite-horizon case. There are several open questions, in particular the complexity of approximate optimization for this class of stochastic controllers. The related literature addresses only the case of deterministic controllers (Lusena et al., 2001).

Acknowledgments

The first author would like to thank Constantinos Daskalakis, Michael Tsatsomeros, John Tsitsiklis, and Steve Vavasis for helpful discussions.

References

- Al-Khayyal, F. A. and Falk, J. E. (1983). Jointly constrained biconvex programming. *Mathematics of Operations Research*, 8(2):273–286.
- Allender, E., Bürgisser, P., Kjeldgaard-Pedersen, J., and Miltersen, P. B. (2009). On the complexity of numerical analysis. *SIAM J. Comput.*, 38(5):1987–2006.
- Amato, C., Bernstein, D. S., and Zilberstein, S. (2007). Solving POMDPs using quadratically constrained linear programs. In *Proc. 20th Int. Joint Conf. on Artificial Intelligence*, Hyderabad, India.

- Bernstein, D. S., Hansen, E. A., and Zilberstein, S. (2005). Bounded policy iteration for decentralized POMDPs. In *Proc. 19th Int. Joint Conf. on Artificial Intelligence*, Edinburgh, Scotland.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge, UK.
- Canny, J. F. (1988). Some algebraic and geometric computations in PSPACE. In *ACM Symposium on Theory of Computing*, pages 460–467.
- Chrisman, L. (1992). Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proc. 10th National Conf. on Artificial Intelligence*, San Jose, CA.
- Etessami, K. and Yannakakis, M. (2010). On the complexity of Nash equilibria and other fixed points. *SIAM Journal on Computing*, 39(6):2531–2597.
- Garey, M. R., Graham, R. L., and Johnson, D. S. (1976). Some NP-complete geometric problems. In *ACM Symposium on Theory of Computing*.
- Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.
- Hansen, E. (1998). Solving POMDPs by searching in policy space. In *Proc. 14th Int. Conf. on Uncertainty in Artificial Intelligence*, Madison, Wisconsin, USA.
- Hastings, N. A. J. and Sadjadi, D. (1979). Markov programming with policy constraints. *European Journal of Operations Research*, 3:253–255.
- Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134.
- Littman, M. L. (1994). Memoryless policies: Theoretical limitations and practical results. In *Proc. 3rd Int. Conf. on Simulation of Adaptive Behavior*, Brighton, England.
- Littman, M. L., Goldsmith, J., and Mundhenk, M. (1998). The computational complexity of probabilistic planning. *Journal of Artificial Intelligence Research*, 9:1–36.
- Lusena, C., Goldsmith, J., and Mundhenk, M. (2001). Nonapproximability results for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 14:2001.
- Madani, O., Hanks, S., and Condon, A. (1999). On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems. In *Proc. 16th National Conf. on Artificial Intelligence*.
- Meuleau, N., Kim, K., Kaelbling, L., and Cassandra, A. (1999). Solving POMDPs by searching the space of finite policies. In *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, Stockholm, Sweden.
- Motzkin, T. S. and Straus, E. G. (1965). Maxima for graphs and a new proof of a theorem of Turán. *Canadian Journal of Mathematics*, 17:533–540.

- Mundhenk, M., Goldsmith, J., Lusena, C., and Allender, E. (2000). Complexity of finite-horizon Markov decision process problems. *Journal of ACM*, 47:681–720.
- Papadimitriou, C. H. and Tsitsiklis, J. N. (1987). The complexity of Markov decision processes. *Mathematics of operations research*, 12(3):441–450.
- Platzman, L. K. (1981). A feasible computational approach to infinite-horizon partially-observed Markov decision problems. Technical report, School of Industrial and Systems Engineering, Georgia Institute of Technology. J-81-2.
- Poupart, P. and Boutilier, C. (2004). Bounded finite state controllers. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA. MIT Press.
- Puterman, M. (1994). *Markov decision processes : Discrete stochastic dynamic programming*. John Wiley & Sons, New York.
- Serin, Y. and Kulkarni, V. G. (2005). Markov decision processes under observability constraints. *Mathematical Methods of Operations Research*, 61:311–328.
- Singh, S. P., Jaakkola, T., and Jordan, M. I. (1994). Learning without state-estimation in partially observable Markovian decision processes. In *Proc. 11th Int. Conf. on Machine Learning*, San Francisco, CA.
- Smith, J. L. (1971). Markov decisions on a partitioned state space. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-1, pages 55–60.
- Sondik, E. J. (1971). *The optimal control of partially observable Markov decision processes*. PhD thesis, Stanford University.
- Toussaint, M., Storkey, A., and Harmeling, S. (2011). Expectation-Maximization methods for solving (PO)MDPs and optimal control problems. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models*. Cambridge University Press.